Introducing the MindSet benchmark for comparing DNNs to human vision

Valerio Biscione (Valerio.biscione@gmail.com)
School of Psychological Science, 12a Priory Road,
Bristol, BS8-1TU

Don Yin (wa18873@alumni.bristol.ac.uk)
School of Psychological Science, 12a Priory Road,
Bristol, BS8-1TU

Gaurav Malhotra (Gaurav.malhotra@bristol.ac.uk)
School of Psychological Science, 12a Priory Road,
Bristol, BS8-1TU

Marin Dujmović (marin.dujmovic@bristol.ac.uk)
School of Psychological Science, 12a Priory Road,
Bristol, BS8-1TU

Milton L. Montero (m.lleramontero@bristol.ac.uk)
School of Psychological Science, 12a Priory Road,
Bristol, BS8-1TU

Guillermo Puebla (guillermo.puebla@bristol.ac.uk)
National Center for Artificial Intelligence,
Vicuña Mackenna 4860, Macul, Chile

Federico Adolfi (fedeadolfi@gmail.com)
Ernst Strüngmann Institute (ESI) for Neuroscience
in Cooperation with Max Planck Society, Germany

Christian Tsvetkov (christian.tsvetkov@bristol.ac.uk)
School of Psychological Science, 12a Priory Road,
Bristol, BS8-1TU

Rachel F. Heaton (mflood2@illinois.edu)
Department of Psychology, University of Illinois
Champaign, IL 61820

John E. Hummel (jehummel@illinois.edu)
Department of Psychology, University of Illinois
Champaign, IL 61820

Benjamin D. Evans (b.d.evants@sussex.ac.uk)
School of Engineering and Informatics
University of Sussex

Jeffrey S. Bowers (j.bowers@bristol.ac.uk)
School of Psychological Science, 12a Priory Road,
Bristol, BS8-1TU

Abstract:

We describe the *MindSet* benchmark designed to facilitate the testing of DNNs against controlled experiments reported in psychology. MindSet will focus on a range of low-, middle-, and high-level visual findings that provide important constraints for theory, provide the materials for testing DNNs, and provide an example of how to assess a DNN on each experiment using a ResNet152 pretrained on ImageNet. The goal is not to evaluate how well ResNet152 accounts for human vision, but rather, encourage researchers to assess how well various DNNs account for a range of key human visual phenomena.

Keywords: Deep Neural Networks; Benchmarks; Vision

Introduction

Deep neural networks (DNNs) provide the best solution to identifying naturalistic images of objects short of biological vision, and many researchers claim that DNNs are the best current models of human visual object processing. The main evidence for this claim comes from the finding that DNNs perform best on various behavioral and brain benchmark tests such as Brain-Score (Schrimpf et al., 2020). The general assumption is that the better a model does at predicting data from benchmark datasets the more biologically plausible the model is.

Bowers et al. (in press) have recently challenged the evidence taken to support DNN-human similarities, for two reasons. First, the experiments included in current benchmarks are treated as observational studies. That is, the data are not organized into any conditions, and consequently, the models are not predicting the impact of any systematic manipulation of images to test specific hypotheses about how vision works. This is a problem as DNNs may identify objects and predict benchmark datasets based on confounds in images that humans ignore (Dujmovic et al., 2022). Second, the same models that perform well on behavioral and brain benchmarks account for almost no findings from vision research in psychology (Bowers et al., in press). Part of the reason for this is that most researchers have not attempted to account for findings reported in psychology, and when these models are explicitly tested against such findings, they generally fail. Furthermore, in cases in which researchers do report that DNNs capture key findings, it is often the case that the conclusions do not stand up when the model is subjected to more severe tests of falsification.

To address both these limitations we are developing a new benchmark called *MindSet* composed of stimuli and methods that make it easy to carry out classic controlled experiments on DNNs. Testing DNNs on controlled experiments that manipulate independent variables makes it less likely that models can predict the data based on confounds, and the inclusion of a range of tests of key properties of human vision provides a more severe test of the correspondences between DNNs and humans.

We have selected studies based on their theoretical importance to theories of vision and our ability to test a standard DNN on the corresponding finding. For each benchmark experiment we: (1) Describe the finding and how it is relevant to theory; (2) Provide the stimuli or the script we used to generate the stimuli for our experiments; and (3) Test a standard feedforward CNN (ResNet152) pretrained on ImageNet after freezing the weights. Human performance is compared to the model output or the outputs of decoders that assess the information encoded in intermediate layers of the DNN.

It is important to emphasize that the goal is not to show how well ResNet152 performs on all the tests. Rather the goal is to make researchers (1) aware of key experiments in psychology; (2) provide researchers with the stimuli needed to assess their model on the experiment; and (3) suggest a method for assessing how well a DNN accounts for the human results. If researchers identify better ways to assess a model against the experimental results we will incorporate these tests to iteratively improve MindSet.

Overview of the MindSet experiments

Here is a list of empirical phenomena that will be included in MindSet organized according to type of visual processing.

Low-level vision: Experiments assess Weber's Law effects, contrast-sensitivity functions, size constancies and associated illusions including the Ponzo illusion; lightness constancies and associated illusions; size contrast effects, including Ebbinghaus and Jastrow illusions.

Gestalt organization: Experiments assess whether contours can be extracted based on texture contrasts and good continuation; whether Gestalt effects camouflage embedded figures; whether symmetry is encoded; whether emergent features are encoded.

Representations of shape: Experiments assess the role of texture vs. shape in object identification; sensitivity to non-accidental features; whether objects are encoded by their parts, relations between parts, or global shape; whether 3D structure in encoded.

High-level vision and reasoning: Experiments assess how robust object recognition is to various forms of noise and distortions such as elastic deformation, color inversion, and occlusion; invariance to rotation in depth and picture plane; and the capacity to make same/different judgments.

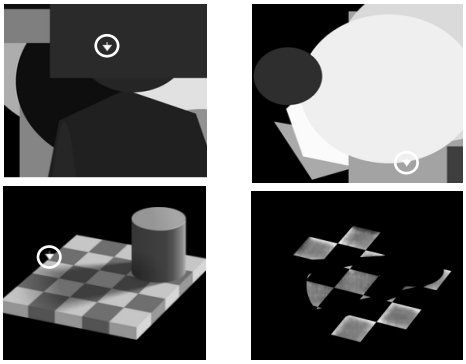
Results

We have created the dataset for approximately half of the task listed above and designed for each one a methodology to compare DNNs to human performance. Given space constraints we have only provided two illustrative findings below.

Lightness constancy

We trained six decoders at progressively deeper stages of ResNet152 to decode the lightness of a single pixel marked by an arrow taken from meaningless grayscale image (see Figure 1 top row for two examples of grayscale images and the arrow – marked by a circle to make clearer -- that indicates the location of the pixel to decode). After training the decoders to output lightness values of a pixel marked by the arrow, we presented the DNN with the classic checker shadow image (Adelson, 1995) (Figure 1 bottom left). We pointed the arrow at all pixel locations of critical tiles in the shadow images and depict the lightness predictions of one of the decoders (Figure 1 bottom right), although similar findings were obtained for all decoders.

Figure 1

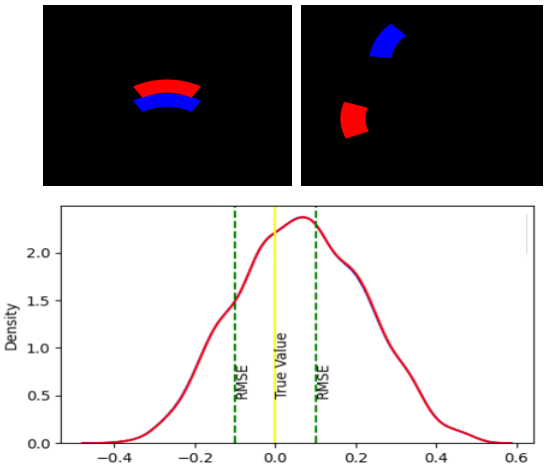


As can be seen in the right bottom panel, the model predicted similar lightness values for all the tiles; that it, the DNN was not affected by the checkerboard illusion.

Size contrast effect

In the Jastrow illusion, the relative size of two curved segments is misperceived when one is placed on top of the other. For example, in the top left panel of Figure 2, the blue disk looks larger than the red one. We trained six decoders at progressively deeper stages of ResNet152 to output the size of varying red and blue segments when the segments were not in configuration that generate the illusion (top right panel), and then applied the decoders when the disks were presented in the relevant configuration. Bottom figure shows the output from a single decoder (similar result obtained across decoders) when the two disks were the same size. The density plot depicts the relative size of the top figure compared to bottom figure in terms of percentage change from the true value – such that a human illusion would take on a negative value (with top figure appearing smaller). Vertical checked bars indicate significant differences from the true value. ResNet152 is not subject to this illusion.

Figure 2



Conclusion

When making claims regarding DNN-human similarities it is important to assess how well DNNs capture key experimental results that manipulate independent variables that test specific hypotheses. This allows researchers to rule out confounds and make causal claims about how a natural system works. The MindSet benchmark includes key findings from psychology and shows how those effects can be easily assessed in controlled experiments using ResNet152. The stimuli or the scripts for generating stimuli will be provided to facilitate testing other DNNs on these experiments. We hope MindSet will redirect efforts to testing DNNs on specific controlled experiments that provide important constraints on human vision, rather than compete on observational datasets to achieve an overall top score. We think this is a more promising approach to evaluating DNN-human similarities and building better models of human vision.

References

Adelson, E. (1995). *Checkershadow illusion*. Retrieved from http://persci.mit.edu/gallery/checkershadow

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., ... & Blything, R. (in press). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*.

Dujmović, M., Bowers, J.S., Adolfi, F., & Malhotra, G. (2022). The pitfalls of measuring representational similarity using representational similarity analysis. *arXiv* https://www.biorxiv.org/content/10.1101/2022.04.05.487135v1

Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., & DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron, 11*, 413-423